Running Multiple Small-Rank MPI Jobs with GNU Parallel and PBS Job Arrays

If you have an MPI application with a small number of ranks and you need to run it many times, each with a different input parameter, consider using both the GNU Parallel tool and the PBS Job Arrays feature, if it meets these criteria:

- The number of MPI ranks is equal to or smaller than the number of physical cores in a node.
- There is enough memory in the node for multiple MPI jobs.

Example

Suppose you have a 5-rank MPI application called **`hello_mpi_processor`,** and you need to run it 400 times with 400 different input parameters (represented by the files **`in_1, in_2, ..,in_400`** in the example). The following PBS script will divide these 400 runs into 10 PBS array sub-jobs. In each sub-job, one Ivy Bridge node is used to fit four of the 5-rank runs simultaneously. The GNU Parallel tool is used to schedule 40 of these runs in each sub-job.

runscript

```
#PBS -S /bin/csh
#PBS -l select=1:ncpus=20:model=ivy
#PBS -l walltime=00:10:00
#PBS -J 1-400:40
#PBS -j oe

cd $PBS_O_WORKDIR

set begin=$PBS_ARRAY_INDEX
set end=`expr $PBS_ARRAY_INDEX + 39`

seq $begin $end | parallel -j 4 -u ./my_hello.scr {}"
```

my_hello.scr

The runscript, shown above, refers to the following script which contains all the tricky details:

```
#!/bin/csh -f

# need to enable module command and load modules
source ~/.cshrc
module load mpi-hpe/mpt

set case=$1
set nprocs=5

#need to create PBS_NODEFILE to run mpiexec
echo `hostname` >>! nodefile_$case
echo `hostname` >>! nodefile_$case
echo `hostname` >>! nodefile_$case
echo `hostname` >>! nodefile_$case
echo `hostname` >>! nodefile_$case
setenv PBS_NODEFILE `pwd`/nodefile_$case

# need to disable pinning to avoid having multiple processes run
# on the same set of CPUs
setenv MPI_DSM_DISTRIBUTE 0

date                                              >  out_$case
echo "running case $case on $nprocs processors"   >>  out_$case
```

```
mpiexec -np $nprocs ./hello_mpi_processor in_$case >>  out_$case
date                                              >>  out_$case

# remove nodefile at end of run
\rm nodefile_$case
```

Job Submission

```
pfe% qsub runscript
1468444[].pbspl1.nas.nasa.gov
```

Output

At the end of the run, you should have files **`out_[1-400]`** and **`runscript.o1468444.[1,41,81,..,361]`**.

The following example shows sample contents of one of the out* files:

```
Wed Mar 15 16:05:12 PDT 2017
running case 42 on 5 processors
...
<output generated by hello_mpi_processor>
...
Wed Mar 15 16:05:13 PDT 2017
```

Article ID: 545
Last updated: 21 Nov, 2022
Revision: 13
Optimizing/Troubleshooting -> PBS Jobs -> Running Multiple Small-Rank MPI Jobs with GNU Parallel and PBS Job Arrays
https://www.nas.nasa.gov/hecc/support/kb/entry/545/